УДК 004.89 doi:10.20998/2413-4295.2020.02.07

АНАЛІЗ ДАНИХ ТА МАШИННЕ НАВЧАННЯ НА ОСНОВІ ДАНИХ ЛАБОРАТОРІЇ ЦЕРН

В. В. ГИГИНЯК*, *, А. О. ХЛЕВНИЙ

кафедра технологій управління, Київський національний університет імені Тараса Шевченка, м. Київ, УКРАЇНА
**e-mail: hyhyniak.viktor@email.com*

АНОТАЦІЯ *У даній роботі проведено аналіз даних, застосовано та порівняно між собою ряд методів машинного навчання до одного із найбільш важливих за своїм впливом та значенням відкритих датасетів організації ЦЕРН, розміщених на CERN Open Data Portal, який пов'язаний із відкриттям бозону Хіггса. Завдання полягало у вирішенні проблеми бінарної класифікації та розподіленні спостережень на ті, що свідчать про сигнал розпаду частинки та фонові. На першому етапі було проаналізовано вхідні дані, проведено аналіз відсутніх значень. Було відзначено залежність факту відсутності більшості змінних від однієї характеристичної, а також перевірено чи впливає наявність/відсутність на приналежність спостережень до сигналу. Для оцінки та отримання початкових результатів про вплив змінних на результат було розраховано матриці кореляцій. Далі застосовано більш точний та надійний метод розрахунку Predictive Power Score, який є новим та перспективним підходом до визначення залежностей, а саме передбачувальних властивостей змінних. Для подальшого застосування підходів машинного навчання датасет було оброблено та очищено, виявлено та закодовано категоріальні змінні за підходом «one-hot encoding», а також проведено заміну відсутніх значень на розраховані середні по датасету. Після підготовки вхідних даних їх було використано для навчання та валідації ряду моделей. Оскільки проблема полягала в вирішенні питання бінарної класифікації, то до розглянутих моделей ввійшли найбільш поширені методи класифікації, такі як: Decision Tree, Logistic Regression, Bagging, Random Forest, K-Nearest Neighbours, Gradient Bossting, XGB, SVM. До кожного з методів було застосовано пошук гіперпараметрів із використанням 2-фолдної крос-валідації. Серед метрик для оцінки якості та продуктивності моделей було обрано метрики акуратності, точності, чутливості, F-значення та AUC, остання з них була вирішальною, оскільки найбільше підходила до вимог та особливостей класифікації. Найкращими себе показали K-Nearest Neighbours та методи, що базуються на побудові ансамблів із простих класифікаторів, а саме дерев рішень. Також було проведено навчання та валідація моделей на базі нейронних мереж, які хоч і показали досить високі результати, однак через проблематику з перенавчанням виявилися дещо гіршими за методи на основі побудови ансамблів. Найвищі значення спостерігались для Gradient Boosting та XGB, а так як останній є схожим за принципом до першого, але має ряд переваг по швидкості, надійності та продуктивності, то було обрано зупинитися саме на ньому. Після наступного етапу вдосконалення вхідних параметрів моделі, було досягнуто збільшення значень метрик та отримано високі показники передбачувальної здатності. Оскільки XGB базується на побудові ансамблів із простіших предикторів (а в даному випадку дерев рішень), то це дозволило отримати наочне уявлення про алгоритм передбачення. Таким чином наступним кроком було проведено візуалізацію роботи отриманої моделі у вигляді побудови зведеного дерева рішень, а також розраховано F-значення важливості змінних. Отримані результати дозволили провести аналіз впливу кожної із змінних на передбачення сигналу, а також порівняти їх із теоретичними відомостями. Було відмічено більший вплив змінних, отриманих вченими методами розрахунку із теоретичних формул в порівнянні із вхідними змінним, які відповідали неопрацьованим значенням детекторів. Таким чином в результаті роботи було проаналізовано різні підходи та методи машинного навчання, встановлено, що найбільш продуктивними та при цьому легкими в інтерпретації результатів є моделі на базі ансамблю дерев рішень, а також отриманий алгоритм для роботи з експериментальними даними, їх аналізом та використанням у методах машинного навчання.*

Ключові слова: *аналіз даних; обробка даних; експериментальні дані; машинне навчання; бінарна класифікація; градієнтний бустинг*

DATA ANALYSIS AND MACHINE LEARNING ON THE CERN DATA

V. HYHYNIAK, A. KHLEVNYI

Department of Management Technologies, Taras Shevchenko National University of Kyiv, Kyiv, UKRAINE

ABSTRACT *The data from one of the open datasets of CERN, applying and comparing a number of machine learning methods were analyzed. The dataset is hosted on the CERN Open Data Portal and is associated with the discovery of the Higgs boson. It is considered to be one of the most challenging in terms of impact and importance. First of all, the task was to solve the problem of binary classification and the division of observations into the records that indicate the signal of particle decay and background. At the first stage, the input data were analyzed, and the missing values were processed too. The fact of the dependence of most variables on the absence of one characteristic was noted, and it was checked whether the presence/absence affects the affiliation of observations to the signal. Correlation matrices were calculated to evaluate and obtain initial results on the influence of variables on the output. Secondly, a more accurate and reliable method of calculating the Predictive Power Score was used, which is a new and promising approach to determine the dependencies, namely the predictive properties of variables. For further application of machine*

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

learning approaches, the dataset was processed and cleaned, categorical variables were identified and coded according to the "one-hot encoding" approach, and the missing values were replaced with the calculated averages of the dataset. After preparing the input data, they were used for training and validation of a number of models. Since the task was to solve the problem of binary classification, the considered models included the most common classification methods, such as: Decision Tree, Logistic Regression, Bagging, Random Forest, K-Nearest Neighbors, Gradient Boosting, XGB, SVM. The search for hyperparameters using 2-fold cross-validation was applied to each of the methods. Among the metrics for assessing the quality and performance of the models, metrics of accuracy, precision, sensitivity, F-value and AUC were chosen, the latter of which was crucial because it best suited the requirements and features of the classification. K-Nearest Neighbors and methods based on building ensembles from simple classifiers, namely decision trees, proved to be the best. Models based on neural networks were also suggested and validated, although they showed quite good results, due to the problem of overfitting turned out to be slightly worse than the methods based on the construction of ensembles. The highest values were observed for Gradient Boosting and XGB, and since the latter is similar in principle to the first one, but has a number of advantages in speed, reliability and performance, it was chosen to focus on. After the next stage of improving the input parameters of the model, an increase in the values of metrics was achieved and high indicators of predictability were obtained. Since XGB is based on building ensembles from simple predictors (and in this case decision trees), this allowed us to get a clear idea of the prediction algorithm. Thus, the next step was to visualize the work of the obtained model in the form of constructing a consolidated decision tree, and also calculate the F-values of the importance of variables. The obtained results allowed us to analyze the influence of each of the variables on the prediction of the signal, as well as to compare them with theoretical information. A greater influence of variables was obtained by scientific methods of calculation from theoretical formulas in comparison with the input variables, which corresponded to the raw values of the detectors. Thus, as a result of the work different approaches and methods of machine learning were analyzed, it was found that the most productive and easy to interpret the results are models based on the ensemble of decision trees, and the algorithm for working with experimental data, their analysis and use in methods of machine learning was established.
***Keywords**: data analysis; data processing; experimental data; machine learning; binary classification; gradient boosting*

Вступ

Роль та вплив науки про дані зростає з кожним роком, аналіз даних та машинне навчання набирають все більшої ваги як невід'ємна частина робочого процесу організацій, шо оперують значними обсягами інформації. До однієї з таких належить і відома у всьому світі ЦЕРН - Європейська організація з ядерних досліджень, яка одночасно є найбільшою в світі лабораторією з фізики елементарних частинок. Проведення експериментів завжди супроводжується генерацією великих об'ємів даних (наприклад, за 2016 рік було згенеровано понад 49 Петабайтів даних [1]), що містять в собі важливу інформацію, яка, однак, потребує знань вчених для її виокремлення.

У зв'язку зі зростаючим інтересом до даних, отриманих в ЦЕРН, було створено CERN Open Data portal [2], на якому знаходиться інформація призначена для навчальних та дослідницьких цілей. На порталі зібрані відкриті дані з різних LHC експериментів: ALICE, ATLAS, CMS та LHCb. Серед найбільш цікавих зі сторони Data Science є підготовані збірки / набори даних для аналізу даних та машинного навчання. В даній роботі було обрано один із найбільш важливих за своїм впливом та значенням датасет, пов'язаний із відкриттям бозону Хіггса. Розглянутий датасет побудований на основі змодельованих за допомогою офіційного генератора-моделі ATLAS full-detector подій "Higgs to tautau", змішаних з різним фоном [3]. Отримані з експериментів дані необхідно класифікувати та перевірити на відповідність подій розпадам бозонів Хіггса. Для ідентифікації сигналу (каналу, або події), що відповідає розпаду бозону, використовуються спеціальні алгоритми, побудовані дослідниками на базі фізичних знань. Але все більшої уваги привертають підходи до класифікації, отримані за допомогою методів машинного навчання. Вони є досить перспективними у напрямку покращення існуючих підходів та алгоритмів. З цією метою організація надала дані для можливості створення нових моделей іншими науковцями та фахівцями з data science.

Мета роботи

В даній роботі ставилося на меті провести аналіз вхідних даних, зробити їх обробку та підготовку до подальшого використання в моделях, а також розробити рішення для поставленої проблеми бінарної класифікації методами машинного навчання [4], а саме: розробити, навчити та порівняти між собою ефективність моделей, визначити найбільш вдалі та доцільні, забезпечити їх легке використання для автоматизації ідентифікації сигналів з даних, які отримуються на експериментах. Отриманий підхід до автоматичного прийняття рішень про класифікацію сигналів, допоможе та пришвидшить роботу під час аналізу отриманих даних з експерименту в середовищі Python [5]. Також отримання алгоритму роботи з даними, розробки та валідації моделей може бути застосований до інших наборів вхідних даних.

Викладення основного матеріалу

На першому етапі роботи було проаналізовано вхідні дані, проведено аналіз відсутніх значень [6]. Було відзначено залежність факту відсутності більшості змінних від однієї характеристичної, а також перевірено чи впливає наявність/відсутність на приналежність спостережень до сигналу. Для оцінки та отримання початкових результатів про вплив змінних на результат було розраховано матриці кореляцій. Далі застосовано більш точний та надійний

метод розрахунку Predictive Power Score [7], який є новим та перспективним підходом до визначення залежностей, а саме передбачуваних властивостей змінних. Для подальшого застосування підходів машинного навчання датасет було оброблено та очищено, виявлено та закодовано категоріальні змінні за підходом «one-hot encoding», а також проведено заміну відсутніх значень на розраховані середні по датасету. Після підготовки вхідних даних їх було використано для навчання та валідації ряду моделей. Оскільки проблема полягала в вирішенні питання бінарної класифікації, то до розглянутих моделей ввійшли найбільш поширені методи класифікації, такі як: Decision Tree, Logistic Regression, Bagging, Random Forest, K-Nearest Neighbours, Gradient Bossting, XGB, SVM. До кожного з методів було застосовано пошук гіперпараметрів з використанням 2-фолдної крос-валідації. Серед метрик для оцінки якості та продуктивності моделей було обрано метрики акуратності, точності, чутливості, F-значення та AUC, остання з них була вирішальною, оскільки найбільше підходила до вимог та особливостей класифікації. Далі було обрано як базовий підхід реалізацію методу Gradient Boosting [8] на базі XGB, які зарекомендували себе як одні з найкращих для завдань класифікації [9] . Після наступного етапу вдосконалення вхідних параметрів моделі, було досягнуто збільшення значень метрик та отримано високі показники передбачувальної здатності. Оскільки XGB базується на побудові ансамблів із простіших предикторів (а в даному випадку дерев рішень), то це дозволило отримати наочне уявлення про алгоритм передбачення. Таким чином наступним кроком було проведено візуалізацію роботи отриманої моделі у вигляді побудови зведеного дерева рішень, а також розраховано F-значення важливості змінних. Отримані результати дозволили провести аналіз впливу кожної із змінних на передбачення сигналу, а також порівняти їх із теоретичними відомостями.

Обговорення результатів

Проведений аналіз відсутніх значень показав наявність великої кількості відсутніх значень (рис. 1) та їх залежність від значень однієї факторної змінної. Після розгляду розподілів значень було прийнято рішення про заповнення відсутніх значень методам розрахунку середніх показників по змінним, що забезпечило можливість використання ряду методів машинного навчання.

Після навчання та валідації моделей було отримано ряд їх характеристик, що наведені на рис. 2.

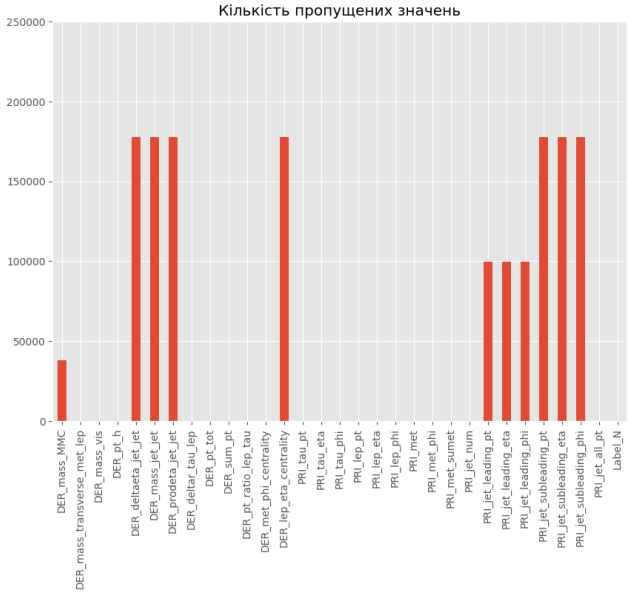

Рис. 1 – Кількість відсутніх значень серед змінних

Classifier	Accuracy	Precision	Recall	FMeasure	AUC
GaussianNB	0.68204	0.546211	0.435912	0.484868	0.761160
LogisticRegression	0.74952	0.675280	0.532837	0.595661	0.815203
KNeighborsClassifier	0.80190	0.740824	0.653073	0.694186	0.868067
DecisionTreeClassifier	0.81890	0.763739	0.677228	0.717886	0.880569
RandomForestClassifier	0.82294	0.792401	0.656752	0.718228	0.889015
SVC	0.83064	0.779958	0.699118	0.737329	0.893542
BaggingClassifier	0.83554	0.790000	0.711381	0.748632	0.899080
XGBClassifier	0.83400	0.783139	0.715871	0.747996	0.904973
GradientBoostingClassifier	0.84154	0.795140	0.725754	0.758864	0.910165

Рис. 2 – Результати валідації різних методів машинного навчання

Найкращими себе показали K-Nearest Neighbours та методи, що базуються на побудові ансамблів із простих класифікаторів, а саме дерев рішень. Також було проведено навчання та валідація моделей на базі нейронних мереж, які хоч і показали досить високі результати, однак через проблематику з перенавчанням виявилися дещо гіршими за методи на основі побудови ансамблів. Найвищі значення спостерігались для Gradient Boosting та XGB, а так як останній є схожим за принципом до першого, але має ряд переваг по швидкості, надійності та продуктивності, то було обрано зупинитися саме на ньому. Оскільки XGB базується на побудові ансамблів із простіших предикторів (а в даному випадку дерев рішень), то це дозволило отримати наочне уявлення про алгоритм передбачення. Таким чином наступним кроком було проведено візуалізацію роботи отриманої моделі у вигляді побудови зведеного дерева рішень, а також розраховано F-значення важливості змінних [10]. Отримані результати дозволили провести аналіз впливу кожної із змінних на передбачення сигналу, а також

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

порівняти їх із теоретичними відомостями. Було відмічено більший вплив змінних, отриманих вченими методами розрахунку із теоретичних формул в порівнянні із вхідними змінним, які відповідали неопрацьованим значенням детекторів

Висновки

В результаті роботи було проаналізовано різні підходи та методи машинного навчання, встановлено, що найбільш продуктивними та при цьому легкими в інтерпретації результатів є моделі на базі ансамблю дерев рішень. Застосування нейронних мереж є також високоефективним, однак в порівнянні з ансамблями дерев рішень не мають настільки високих показників по надійності передбачень та можливості простої інтерпретації отриманої моделі та визначення впливу кожної окремої змінної на результат. Також було отримано алгоритм для роботи з експериментальними даними, їх аналізом, обробкою та використанням в методах машинного навчання.

Список літератури

1. *Annual Report 2016*. CERN. Retrieved 22 September 2017. URL: https://cds.cern.ch/record/2270805/files/486-1611-1-SM.pdf (дата звернення 18.01.2020).
2. *CERN Open Data Portal*. URL: http://opendata.cern.ch/ (дата звернення: 05.05.2020).
3. ATLAS collaboration. *Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014*. CERN Open Data Portal. 2014. doi:10.7483/opendata.atlas.zbp2.m5t8.
4. Alpaydin Ethem. *Introduction to Machine Learning*. 2010. MIT Press. p. 9.
5. Fortune Nathanael Alexander. *A Short Guide to Using Python For Data Analysis In Experimental Physics*. 2018. Physics: Faculty Publications, Smith College, Northampton, MA. URL: https://scholarworks.smith.edu/phy_facpubs/30 (дата звернення 18.01.2020).
6. Salgado C. M., Azevedo C., Proença H., Vieira S. M. *Missing Data. In: Secondary Analysis of Electronic Health Records*. 2016. Springer, Cham. P. 143–162.
7. RIP correlation. Introducing the Predictive Power Score : веб-сайт. URL: https://8080labs.com/blog/posts/rip-correlation-introducing-the-predictive-power-score-pps/ (дата звернення: 03.05.2020).
8. Gérard Biau, Benoît Cadre. *Optimization by gradient boosting*. 2017. URL: ffhal-01562618 (дата звернення 18.01.2020).
9. Changming Zhao, Dongrui Wu, Jian Huang, Ye Yuan, Hai-Tao Zhang. *BoostTree and BoostForest for Ensemble Learning*. URL: arXiv:2003.09737 (дата звернення 18.01.2020).
10. *Feature Importance and Feature Selection With XGBoost in Python*. URL: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/ (дата звернення: 03.05.2020).

References (transliterated)

1. *Annual Report 2016*. CERN. Retrieved 22 September2017. Available at: https://cds.cern.ch/record/2270805/files/486-1611-1-SM.pdf (accessed 18.01.2020).
2. *CERN Open Data Portal*. Available at: http://opendata.cern.ch/ (accessed: 05.05.2020).
3. ATLAS collaboration. Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal, 2014, doi:10.7483/opendata.atlas.zbp2.m5t8.
4. Alpaydin Ethem. *Introduction to Machine Learning*. MIT Press, 2010, p. 9.
5. Fortune Nathanael Alexander. *A Short Guide to Using Python For Data Analysis In Experimental Physics*. 2018. Physics: Faculty Publications, Smith College, Northampton, MA. Available at: https://scholarworks.smith.edu/phy_facpubs/30 (accessed 18.01.2020).
6. Salgado C.M., Azevedo C., Proença H., Vieira S.M. Missing Data. In: Secondary Analysis of Electronic Health Records. Springer, 2016, Cham, pp.143–162.
7. RIP correlation. Introducing the Predictive Power Score. Available at: https://8080labs.com/blog/posts/rip-correlation-introducing-the-predictive-power-score-pps/ (accessed: 03.05.2020).
8. Gérard Biau, Benoît Cadre. *Optimization by gradient boosting*. 2017. Available at: ffhal-01562618 (accessed 18.01.2020).
9. Changming Zhao, Dongrui Wu, Jian Huang, Ye Yuan, Hai-Tao Zhang *BoostTree and BoostForest for Ensemble Learning*. Available at: arXiv:2003.09737 (accessed 18.01.2020).
10. *Feature Importance and Feature Selection With XGBoost in Python*. Available at: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/ (accessed: 03.05.2020).

Сведения об авторах (About authors)

Гигиняк Віктор Васильович – магістрант кафедри технологій управління, Київський національний університет імені Тараса Шевченка; м. Київ, Україна; e-mail: hyhyniak.viktor@gmail.com.

Viktor Hyhyniak – Bachelors Degree, pursuing Master Degree, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine; e-mail: hyhyniak.viktor@gmail.com.

Хлевний Андрій Олександрович – кандидат технічних наук, асистент кафедри технологій управління, Київський національний університет імені Тараса Шевченка; м. Київ, Україна; e-mail: andlev@ukr.net.

Andrii Khlevnyi – Candidate of Technical Sciences (Ph. D.), Department of, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine; e-mail: andlev@ukr.net.

Будь ласка, посилайтесь на цю статтю наступним чином:

Гигиняк В. В., Хлевний А. О. Аналіз даних та машинне навчання на основі даних лабораторії ЦЕРН. *Вісник Національного технічного університету «ХПІ». Серія: Нові рішення в сучасних технологіях*. – Харків: НТУ «ХПІ». 2020. № 2 (4). С. 53-57. doi:10.20998/2413-4295.2020.02.07.

Please cite this article as:
Hyhyniak V., Khlevnyi A. Data Analysis and Machine Learning on the CERN data. *Bulletin of the National Technical University "KhPI". Series: New solutions in modern technology*. – Kharkiv: NTU "KhPI", 2020, no. 2 (4), pp. 53–57, doi:10.20998/2413-4295.2020.02.07.

Пожалуйста, ссылайтесь на эту статью следующим образом:
Гигиняк В. В., Хлевной А. А. Анализ данных и машинное обучение на основе данных лаборатории ЦЕРН. *Вестник Национального технического университета «ХПИ». Серия: Новые решения в современных технологиях*. – Харьков: НТУ «ХПИ». 2020. № 2 (4). С. 53-57. doi:10.20998/2413-4295.2020.02.07.

АНОТАЦИЯ *В данной работе был проведен анализ данных, применены и сравнены между собой ряд методов машинного обучения к одному из самых важных по своему влиянию и значению открытых наборов данных организации ЦЕРН, размещенных на CERN Open Data Portal, который связан с открытием бозона Хиггса. Задача состояла в решении проблемы бинарной классификации и распределении наблюдений на те, что свидетельствуют о сигнале распада частицы и фоновые. На первом этапе были проанализированы входные данные, проведен анализ недостающих значений. Было отмечено зависимость факта отсутствия большинства переменных от одной характеристической, а также проверено влияние наличия/отсутствия переменных на принадлежность наблюдений к сигналу. Для оценки и получения начальных результатов о влиянии переменных на результат была рассчитана матрица корреляций. Далее применено более точный и надежный метод расчета Predictive Power Score, который является новым и перспективным подходом к определению зависимостей, а именно к предсказанию свойств переменных. Для дальнейшего применения подходов машинного обучения, данные было обработано и очищено, обнаружено и закодировано категориальные переменные по методу «one-hot encoding», а также проведена замена отсутствующих значений на рассчитанные средние по набору данных. После подготовки исходных данных они были использованы для обучения и валидации ряда моделей. Поскольку проблема заключалась в решении вопроса бинарной классификации, то в число рассматриваемых моделей вошли наиболее распространенные методы классификации, такие как: Decision Tree, Logistic Regression, Bagging, Random Forest, K-Nearest Neighbours, Gradient Bossting, XGB, SVM. К каждому из методов было применено поиск гиперпараметров с использованием 2-фолдной кросс-валидации. Среди метрик для оценки качества и производительности моделей были выбраны метрики аккуратности, точности, чувствительности, F-значения и AUC, последняя из них была решающей, поскольку больше согласовывалась с требованиями и особенностями классификации. Лучшими себя показали K-Nearest Neighbours и методы, основанные на построении ансамблей из простых классификаторов, а именно деревьев решений. Также было проведено обучение и валидация моделей на базе нейронных сетей, которые хотя и показали достаточно высокие результаты, однако через проблематику с переобучением оказались несколько хуже методов на основе построения ансамблей. Высокие значения наблюдались для Gradient Boosting и XGB, а так как последний похож по принципу на первый, но имеет ряд преимуществ по скорости, надежности и производительности, то было решено остановиться именно на нем. После следующего этапа совершенствования входных параметров модели, было достигнуто увеличение значений метрик и получены высокие показатели предсказательной способности. Поскольку XGB базируется на построении ансамблей из простых предикторов (а в данном случае деревьев решений), то это позволило получить наглядное представление об алгоритме предсказания. Таким образом следующим шагом было проведено визуализацию работы полученной модели в виде построения сводного дерева решений, а также рассчитано F-значение важности переменных. Полученные результаты позволили провести анализ влияния каждой из переменных на предсказания сигнала, а также сравнить их с теоретическими сведениями. Было отмечено большее влияние переменных, полученных учеными методами расчета из теоретических формул по сравнению с входными переменными, которые отвечали необработанным значениям детекторов. Таким образом в результате работы были проанализированы различные подходы и методы машинного обучения, установлено, что наиболее продуктивными и при этом легкими в интерпретации результатов есть модели на базе ансамбля деревьев решений, а также был получен алгоритм для работы с экспериментальными данными, их анализом и использованием в методах машинного обучения.*

Ключевые слова: *анализ данных; обработка данных; экспериментальные данные; машинное обучение; бинарная классификация; градиентный бустинг*

Надійшла (received) 16.05.2020